

AD-A230 127

Technical Report 913

DTIC FILE COPY

①

# **A Meta-Analytic Approach for Relating Subjective Workload Assessments with U.S. Army Aircrew Training Manual (ATM) Ratings of Pilot Performance**

**John E. Stewart II and Ronald J. Lofaro**  
U.S. Army Research Institute

**September 1990**

**DTIC**  
**ELECTE**  
**DEC 18 1990**  
**S B D**



**United States Army Research Institute  
for the Behavioral and Social Sciences**

Approved for public release; distribution is unlimited

**90 12 17 097**

# **U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES**

**A Field Operating Agency Under the Jurisdiction  
of the Deputy Chief of Staff for Personnel**

**EDGAR M. JOHNSON**  
**Technical Director**

**JON W. BLADES**  
**COL, IN**  
**Commanding**

---

Technical review by

N. Joan Blackwell  
Charles A. Gainer  
Donald B. Headley  
David R. Hunter

## **NOTICES**

**DISTRIBUTION:** Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

**FINAL DISPOSITION:** This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

**NOTE:** The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

## REPORT DOCUMENTATION PAGE

Form Approved  
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified			1b. RESTRICTIVE MARKINGS --		
2a. SECURITY CLASSIFICATION AUTHORITY --			3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.		
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE --			5. MONITORING ORGANIZATION REPORT NUMBER(S) --		
4. PERFORMING ORGANIZATION REPORT NUMBER(S) ARI Technical Report 913			7a. NAME OF MONITORING ORGANIZATION --		
6a. NAME OF PERFORMING ORGANIZATION U.S. Army Research Institute Aviation R&D Activity	6b. OFFICE SYMBOL (If applicable) PERI-IR	7b. ADDRESS (City, State, and ZIP Code) --			
6c. ADDRESS (City, State, and ZIP Code) Fort Rucker, AL 36362-5354		9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER --			
8a. NAME OF FUNDING / SPONSORING ORGANIZATION U.S. Army Research Institute for the Behavioral and Social Sciences	8b. OFFICE SYMBOL (If applicable) PERI-I	10. SOURCE OF FUNDING NUMBERS			
8c. ADDRESS (City, State, and ZIP Code) 5001 Eisenhower Avenue Alexandria, VA 22333-5600		PROGRAM ELEMENT NO. 62785A	PROJECT NO. 790	TASK NO. 1211	WORK UNIT ACCESSION NO. H02
11. TITLE (Include Security Classification) A Meta-Analytic Approach for Relating Subject Workload Assessments with U.S. Army Aircrew Training Manual (ATM) Ratings of Pilot Performance					
12. PERSONAL AUTHOR(S) Stewart, II, John E.; and Lofaro, Ronald J.					
13a. TYPE OF REPORT Final	13b. TIME COVERED FROM 89/06 TO 89/11	14. DATE OF REPORT (Year, Month, Day) 1990, September		15. PAGE COUNT	
16. SUPPLEMENTARY NOTATION --					
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)		
FIELD	GROUP	SUB-GROUP	Subjective workload assessment Delphi		
05	08		Nominal group methods Aircrew training		
			Aviation safety		
19. ABSTRACT (Continue on reverse if necessary and identify by block number) In 1985 Lofaro, using a modified Delphi technique, had subject matter experts (SMEs) generate estimated ratings of the subjective workload imposed by various Aircrew Training Manual (ATM) tasks for several Army helicopters, including the UH-60 Blackhawk. For each task, ratio-scaled estimates of difficulty and time to perform were derived. This research was performed to determine the validity of the UH-60 ATM estimates by correlating them with instructor pilot (IP) ratings of checkride performance from two other unrelated research projects. The other efforts investigated the decay of ATM task-related skills among Reserve and regular Army aviators. A second phase of this project compared the difficulty ratings of ATM tasks associated with UH-60 accidents over FY 1980-1988 with those not associated with UH-60 accidents. A negative correlation between the modified Delphi weights assigned to ATM tasks and IP ratings on these tasks was hypothesized; the hypothesis was confirmed. Analysis of the UH-60 accident data confirmed the second hypothesis: ATM tasks that were accident-related had significantly higher Delphi weights than ATM tasks not related (Continued)					
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION Unclassified		
22a. NAME OF RESPONSIBLE INDIVIDUAL Charles A. Gainer			22b. TELEPHONE (Include Area Code) (205) 255-4404	22c. OFFICE SYMBOL PERI-IR	

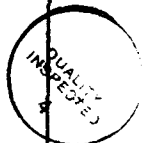
UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

ARI Technical Report 913

19. ABSTRACT (Continued)

to accidents. The report discusses practical applications of the modified Delphi technique, with an emphasis on enhancing aviation safety and improving training effectiveness.



Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

**Technical Report 913**

**A Meta-Analytic Approach for Relating Subjective  
Workload Assessments with U.S. Army Aircrew  
Training Manual (ATM) Ratings  
of Pilot Performance**

**John E. Stewart II and Ronald J. Lofaro**  
U.S. Army Research Institute

**Aviation R&D Activity at Fort Rucker, Alabama**  
**Charles A. Gainer, Chief**

**Systems Research Laboratory**  
**Robin L. Keese, Director**

U.S. Army Research Institute for the Behavioral and Social Sciences  
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel  
Department of the Army

**September 1990**

---

**Army Project Number**  
**2Q162785A790**

**Human Performance Effectiveness**  
**and Simulation**

*Approved for public release; distribution is unlimited.*

## FOREWORD


---

The U.S. Army Research Institute Aviation Research and Development Activity (ARIARDA) provides support enhancing the effectiveness of Army aviator training. One important application of this training research support is to aviation safety. Every operational Army aircraft has an Aircrew Training Manual (ATM) that specifies those tasks necessary for operating the aircraft and how a pilot's performance should be evaluated on each task. The ATM does not, however, provide guidance on the difficulty of the tasks.

The present research effort examined ATM tasks common to two utility helicopters, the UH-1 and the newer UH-60. It involved secondary analysis of data that had been previously collected and analyzed as part of three projects which, though unrelated to each other, were pertinent to the ATM tasks for the utility helicopter mission. The objectives were to examine the relationship between estimated ratings of performance difficulty and time to perform specific ATM tasks for the UH-60 and other variables with relevance to pilot performance and safety. The results indicate that methods used for determining the difficulty of the ATM tasks have validity.

This project was initiated in October 1989 by the Safety Team of ARIARDA at Fort Rucker, Alabama, pursuant to Research Task 1211: Reducing Army Accident Rates in Aviation and Ground Operations. The original modified Delphi analyses, upon which much of the current research is based, were initiated in 1985 as a technical advisory service provided by ARIARDA to the Directorate of Training and Doctrine at Fort Rucker.

The findings of the current research effort suggest a valid means for assessing subjective workload and identifying those ATM tasks aviators are likely to have difficulty performing. The results suggest training interventions that could serve to modify current training standards for these high risk tasks, thereby reducing the probability of aviation accidents.

  
EDGAR M. JOHNSON  
Technical Director

A META-ANALYTIC APPROACH FOR RELATING SUBJECTIVE WORKLOAD  
ASSESSMENTS WITH U.S. ARMY AIRCREW TRAINING MANUAL (ATM) RATINGS  
OF PILOT PERFORMANCE

EXECUTIVE SUMMARY

---

Requirement:

This project was conducted to investigate the validity of subjective workload measures of Aircrew Training Manual (ATM) tasks in relationship to ratings of pilot checkride performance on these tasks.

Procedure:

The subjective workload measures for the UH-60 helicopter, derived through an earlier modified Delphi research project, (Lofaro, 1985) were correlated with instructor pilot (IP) ratings of pilot performance from two other research projects that examined skill decay and reacquisition for ATM tasks. Delphi ratings of ATM tasks associated with UH-60 accidents were also compared to those ratings of tasks that were not associated with accidents for this aircraft.

Findings:

The modified Delphi estimates were found to correlate highly with IP ratings of pilot performance on each of the ATM research projects. Modified Delphi estimates of task difficulty correlated more highly with the criterion IP ratings than did estimates of time to perform. Delphi ratings of difficulty were significantly higher for accident-related ATM tasks than for tasks that were not accident-related.

Utilization of Findings:

The findings demonstrate that the modified Delphi estimates have validity as subjective estimates of pilot workload. The potential exists for their use in determining training standards that could diminish the probability of aviation accidents.

A META-ANALYTIC APPROACH FOR RELATING SUBJECTIVE WORKLOAD  
ASSESSMENTS WITH U.S. ARMY AIRCREW TRAINING MANUAL (ATM) RATINGS  
OF PILOT PERFORMANCE

CONTENTS

---

	Page
INTRODUCTION.....	1
Overview.....	1
Background and History.....	1
Purpose and Rationale.....	6
Hypotheses.....	8
PROCEDURES AND RESULTS.....	8
Overview.....	8
Findings.....	9
DISCUSSION.....	14
Correlations with IP Ratings.....	15
Accident Prevention Usage.....	15
Limitations.....	16
REFERENCES.....	17
APPENDIX A.....	A-1

LIST OF TABLES

Table 1. Aircrew Training Manual (ATM) psychomotor tasks assessed by Wick, et al. (1986).....	4
2. ATM psychomotor tasks assessed by Ruffner & Bickley (1985).....	5
3. Log modified Delphi ratings of difficulty and time to perform for ATM tasks common to Wick, et al. (1986) and Ruffner & Bickley (1985).....	11
4. Aircrew Training Manual (ATM) tasks associated with UH-60 accidents.....	13



A META-ANALYTIC APPROACH FOR RELATING SUBJECTIVE WORKLOAD  
ASSESSMENTS WITH U.S. ARMY AIRCREW TRAINING MANUAL (ATM) RATINGS  
OF PILOT PERFORMANCE

INTRODUCTION

Overview

Each U.S. Army operational helicopter has an Aircrew Training Manual (ATM), which specifies conditions and standards of pilot performance required to operate the aircraft. Each ATM has a reference number and title. The UH-60A "Blackhawk" ATM, or Training Circular 1-212, lists tasks such as Task 1028: "Perform VMC (visual meteorological conditions) approach." It states the conditions, (aircraft and prelanding checks), standards (airspeed and altitude) and presents a brief description of how to perform this task (See Appendix A). In order for a pilot to demonstrate proficiency in an aircraft, he must show satisfactory performance on ATM tasks necessary for piloting the aircraft and selected ATM tasks pertinent to specific missions. The ATM does not provide an exhaustive listing of all UH-60 tasks. Basic aviator tasks are numbered in the 1000 series and special tasks which may be assigned by the unit commander, in the 2000s. Additional unit tasks, which the commander may also assign, are listed as 3000-series tasks, but are not included in the publication.

This report will examine prior research efforts and methodologies which have dealt with U.S. Army ATM tasks. The three research projects discussed in the present report each approached the ATM tasks from differing perspectives and for different purposes. The authors' purpose is to compare the results of these efforts and ascertain how the results can be compared and correlated to yield new insights and to suggest new directions for future research. The title refers to a meta-analytic approach, rather than meta-analysis (Glass, 1976). This was done to denote that, while the present report is in part a summary of other research efforts, and will amass data as part of a comparison of research results, it will not deal with effect sizes per se. Still, it will be more than a narrative review in that the various data will be addressed and re-analyzed, for purposes of exploring the relationship between subject matter expert (SME) ratings of performance difficulty on ATM tasks and other ratings of pilot performance.

Background and History

The Background section to follow will provide the reader with an understanding of the relevant aspects of prior efforts and for the rationale, assumptions, and hypotheses presented later.

Lofaro's modified Delphi approach. In 1985, Lofaro, of the Army Research Institute for the Behavioral and Social Sciences (ARI), devised a highly modified Delphi (Dalkey, 1969) and small-

group-based set of procedures for eliciting SME input and evaluations. He modified the traditional Delphi processes to utilize (a) formal instruction for the participants in group processes, dynamics and methods of consensus, (b) a guided exercise in group consensus followed by evaluation and critique of the group techniques by both group members and a facilitator, (c) a blending, in selected steps of modified Delphi, of anonymous individual ratings with group discussions and consensus (a step-wise procedure based on iterative ratings), (d) use of selected objectives in which the data base for each step in an objective evolved from the preceding steps, and (e) group discussion and consensus as the only rating methods on other selected steps and objectives.

Lofaro conducted three separate two-week workshops using his modified Delphi methodology. Each workshop used 10 SMEs and dealt with a specific U.S. Army helicopter. For the particular helicopter, each ATM task was rated for difficulty to perform as well as actual time to perform for the novice, average and superior Army aviator. Additional work was done on how best to train (in the simulator, aircraft, or some combination of both), as well as the number of iterations needed every six months to maintain proficiency. Finally, some 23 mission profiles were decomposed into all ATM tasks required to complete each mission, evaluated and rank-ordered for difficulty to perform, criticality for mission success and for aircrew safety. A total of 82 performance-related ATM tasks, evaluated in this way, were deemed usable for purposes of the present project in that they corresponded to both the UH-1 and UH-60 ATM tasks. The overlap between these two sets of ATMs is not perfect. For example, one task "takeoff to a hover", which is listed in the UH-1 ATM as Task 2001, does not appear as a separate task in the UH-60 ATM, but is subsumed under Task 1018, "normal takeoff." However, the correspondence between most base UH-1 and UH-60 ATM tasks is high enough to make comparison fairly simple.

A portion of the methodology used by Lofaro in assessing perceived task difficulty was based on the psychophysical method of magnitude estimation (S.S. Stevens, 1971). To establish a ratio scale of difficulty, the ATM tasks were compared to a standard (modulus) low-to-average difficulty task assigned a value of 80. Following the Delphi approach, these comparative estimates of performance difficulty were made independently and anonymously at first, then iterated. This was followed by Lofaro's modification of using group discussion, more iterated ratings, and finally consensus.

The data to be used in the present report are concerned with the difficulty to perform each ATM task, and the time to perform it, for the average aviator. The other data may have some value in future aircrew coordination and simulator-use projects.

The ATM-based decay-reacquisition study of Wick, et al. (1986). Wick, Millard, and Cross (1986) conducted an experiment focusing on the time needed to reacquire ATM-based flying skills. Their sample consisted of 47 experienced reserve aviators (Median= 1260 hr) who had not flown for an average of 7.5 years (range= 1-19). Wick, et al. (1986) looked at the time needed to reacquire flying skills, using proficiency at ATM tasks as a baseline measure. Some 40 ATM tasks (30 psychomotor and 10 procedural) were used to evaluate VMC flight.

Table 1 presents the 30 psychomotor ATM tasks. In the Lofaro project, 25 of these ATM tasks were evaluated via the modified Delphi technique, which imparts a high degree of correspondence across both projects.

Table 1

Aircrew Training Manual (ATM) Psychomotor Tasks Assessed by Wick, et al. (1986)

ATM Task Description	IP Rating
Antitorque malfunction	3.00
Standard autorotation	3.27
Emergency procedures	3.42
IFR recovery procedures	3.50
Low level autorotation	3.57
Hydraulic failure	3.79
Manual throttle operations	3.97
Engine failure (altitude)	4.18
Maximum performance takeoff	4.26
Hover power check	4.31
Steep approach	4.31
Normal approach	4.33
Hovering autorotation	4.33
Shallow approach	4.37
Confined area operations	4.44
Normal takeoff	4.46
Pinnacle & ridgeline operations	4.48
Engine failure (hover)	4.53
Deceleration-acceleration	4.55
Go-around	4.58
High reconnaissance	4.58
Traffic pattern	4.63
Takeoff to hover	4.65
Hovering turn	4.70
Slope operations	4.79
Climb-descents	4.85
Turns	4.85
Hovering flight	4.90
Straight & level flight	4.90
Landing from hover	5.03

Note. Ratings are on a 7-point scale, with 7 being the highest. A rating of 6 means that all ATM standards for a task have been met. These ratings were given on the initial currency flight.

The ATM-based decay-reacquisition study of Ruffner & Bickley (1985). The Ruffner and Bickley (1985) project provides another criterion against which the Delphi ratings can be validated. In this research 79 Army aviators, all UH-1 qualified and current, participated in an ATM skill decay and reacquisition experiment. Ruffner and Bickley's sample consisted of Regular Army staff officers, rather than reserve officers, who had a comparable

number of rotary wing flight hours (Median= 915), and who were not required to fly as part of their duties. These aviators were divided into four groups. Each group flew a different number of iterations of selected ATM flight tasks (see Table 2) in order to ascertain if flight performance skills decayed through lack of practice.

Table 2

ATM Psychomotor Tasks Assessed by Ruffner & Bickley (1985).

ATM Task Description	Checkride	
	Initial	Final
NOE deceleration	7.25	7.25
Engine failure (altitude)	7.50	7.83
Terrain flight takeoff	7.58	8.08
Terrain flight navigation	7.48	8.05
Antitorque malfunction	5.32	6.30
Standard autorotation	6.22	6.59
Terrain flight approach	7.71	8.26
Takeoff to hover	8.04	8.06
Landing from hover	8.03	8.08
Engine failure at hover	7.44	7.55
Confined area ops.	7.49	8.05
Hydraulic failure	7.18	6.89
Normal takeoff	7.90	8.01
Maximum performance takeoff	7.45	7.55
Steep approach	7.47	7.63
Go around	8.00	8.17
Climb-descent	7.88	8.15
Pinnacle-ridgeline operations	7.51	7.76
Straight & level flight	8.19	8.06
Turns	7.87	8.15
Hover power check	8.00	8.06
Traffic pattern flight	7.88	8.09
Hovering flight	8.54	8.23
Acceleration-deceleration	7.92	7.91

Note. These IP ratings employed a 12-point scale; a score of 8 means that all ATM standards were met.

One of these groups flew none of the ATM iterations during the six month period; the others flew either two, four, or six iterations of the selected ATM tasks. No significant difference in the level of psychomotor skills and performance was found for any of these groups, as measured by a pre- and post-experimental

checkride. A closer examination of the data reveals that the majority of ATM tasks used were heavily dependent upon psychomotor skills (e.g.; approaches and hovers) and that procedural (cognitive) ATM skills did indeed show some decay over time for the experimental group with no practice iterations. This latter finding, though informative, is beyond the scope of the present report. It is reported here because of its connection to skill and task analyses as well as to workload analyses.

### Purpose and Rationale

Difficulty and workload. In terms of potential investigations, the most useful data to come out of the modified Delphi project were the difficulty ratings for the ATM tasks. While difficulty does not define all of the complex construct of workload, it nevertheless appears quite pertinent to it. Hart and Bortolussi (1984), for example, found high correlations between pilots' ratings of the effort, stress, and workload. Thus it would seem reasonable to assume that a key determinant of workload is effort; that is to say, the difficulty of the task itself, and how long it must be performed. Both of these factors tie up information processing resources and create situations where errors are likely to occur.

Gopher and Braune (1984) used Stevens' methodology to elicit workload estimates from subjects who performed various perceptual-motor tasks, using a one-dimensional tracking task with a difficulty rating of 100 as the modulus. These workload estimates correlated highly ( $r = .93$ ) with a subjective rating index of task difficulty for each task suggested for this particular study by Wickens. However, correlations with actual performance times on these tasks, though significant, were modest ( $r = .30$ ). The investigators interpreted their findings as supportive of a single-resource model of workload; subjects were able to evaluate all tasks with a single dimension. They were also able to predict dual-task conditions from single-task units with a simple additive model. This was true even though tasks were quite diverse in modalities and mental operations required to perform them. The investigators concluded that they found no evidence that some tasks competed with each other for common resources whereas others did not; the difficulty of the individual tasks was all that seemed to matter. They cautioned, however, that this finding of a single dimension underlying the subjective assessment of workload is limited to the conscious perception of task demands.

Consistent with the rationale for the present research project, one would expect increased task demands to lead to increases in the incidence of errors (see Casali & Wierwille, 1983). Some investigators have gone so far as to state that subjective assessments of task difficulty have inherent validity, in the sense that if one performing a task states that it is difficult or that he or she is overloaded by it, then this must

be true (Moray, et al., 1979). Likewise, a recent study by Vidulich and Tsang (1985), in which two techniques for subjective workload assessment were validated, showed that the more difficult a task was rated to be, the worse the subjects' performance. Consequently, it would be reasonable to suppose that those tasks rated as most difficult should manifest poorer performance measures and more errors than those which are rated as least difficult. Morris and Rouse (1985) point out that whereas high subjective workload should increase the probability of slips and errors occurring, thereby diminishing performance, a case can also be made for extremely low subjective workload having the same effect (underload). For purposes of the present investigation, it is easier to specify those Delphi ratings of ATM tasks which are overloaded than those which are underloaded. Still, the suggestion of a curvilinear relationship is intriguing and invites future inquiry.

These findings strongly suggest that subjective ratings of difficulty, or task demand, by persons familiar with these tasks, can be treated as workload measures. These in turn can be used to predict performance on these same tasks, and to identify potential "problem" tasks that may be excessively difficult for one person to perform.

The initial goal of the researchers was to ascertain if any correlations existed among different means of assessing ATM tasks (e.g. difficulty and time to perform). Since three separate ARI-sponsored projects addressed human performance aspects of ATM tasks, the investigators saw an opportunity to determine if the 1985 modified Delphi ratings could be validated, and whether it had potential as a workload estimation tool.

Further, deterioration of performance on psychomotor tasks should provide a sensitive measure of task difficulty; the pilots in the Ruffner and Bickley and the Wick, et al. projects should perform worse on the more difficult tasks on the initial (baseline) proficiency flight than on those which are less demanding. Thus, the criterion against which the Delphi ratings would be correlated was the performance ratings given by IPs on this flight. These should correlate highly to the extent that the original ratings reflect valid estimates of task demand.

Difficulty and accidents. The U.S. Army Safety Center has recently developed a comprehensive, on line accident reporting system called the Army Safety Management Information System (ASMIS). Of particular interest to the current investigators were the ATM tasks reported by ASMIS as being performed when a given accident occurred. This presented the opportunity to compare the Delphi ATM weights of UH-60 accidents attributed to pilot error with those of ATM tasks which did not appear in the ASMIS reports, for Fiscal Years (FYs) 1980-1988. If the more difficult tasks are the more hazardous, then those ATM tasks associated with accidents should have significantly higher difficulty ratings than those which are not.

## Hypotheses

From the foregoing discussion it would be reasonable to expect that the modified Delphi technique could be used to construct a simple index of relative workload. Proficiency checkride performance ratings could then be used to validate the subjective weights assigned to the Delphi ratings.

Delphi ratings of task difficulty should correlate significantly and negatively with IP ratings of performance on both initial and final checkrides. Likewise, Delphi estimates of time required to perform ATM tasks should correlate positively with the ratings of difficulty for the same tasks. Although it seems reasonable to suppose that estimated time to perform an ATM task should correlate negatively and significantly with IP ratings of performance, it would be difficult to specify in advance the strength of this relationship. While much of the previously-discussed research on subjective workload assessment implies that rated difficulty of a task is highly correlated with ratings of performance on the task, such a case cannot be made with the same confidence for estimates of performance time. It does not necessarily follow, then, that a time-consuming task will inevitably be more difficult than a task with lesser time demands. In fact, one could argue that, in some instances, a task can be difficult because there is not enough time in which to perform it.

Finally, those Delphi difficulty ratings of ATM tasks which are reported by ASMIS should be significantly higher than those which were not reported in conjunction with UH-60 accidents over FYs 1980-1988.

## PROCEDURES AND RESULTS

### Overview

The first step was to construct an index of relative workload from the Delphi data currently available, which could then be used to identify "high-risk" ATM tasks. (High difficulty and high performance time). Concurrent validation of these ratings against measures of proficiency checkride performance should give an indication of how closely the subjective task ratings of one group of IPs correlate with performance ratings by another group.

Recall that two recent ARI-sponsored projects (Wick, Millard & Cross, 1986; Ruffner & Bickley, 1985) sought to evaluate Army training standards and proficiency requirements for the UH-1 helicopter. The modified Delphi ratings of task difficulty were made independently of the ratings of pilot performance, by different raters.



For Wick, et al., a total of 25 ATM tasks were compared which were generic in the sense that they comprised base tasks for the utility helicopter mission, regardless of the type of aircraft; the corresponding number of tasks for Ruffner & Bickley was 24. The Wick, et al. ratings were made on a seven-point scale ranging from one (lowest) to seven (highest). A rating of six was considered passing on any given task; for Ruffner and Bickley, a rating of eight on a 12-point scale was considered passing (all ATM standards for the task were met).

It should be noted that the Lofaro Delphi estimates concerned the UH-60, whereas the Wick, et al. project concerned itself with the UH-1. Both are utility aircraft with overlapping missions; thus the number of common basic ATM tasks is sufficient to allow comparisons. The methodology employed for the present analysis was quite simple and straightforward: Delphi ratings of task difficulty and time to perform were correlated with corresponding IP ratings of initial checkride performance on the two previously-mentioned ARI-sponsored projects, and with final checkride performance as well on the Ruffner and Bickley project.

### Findings

Correlation with Wick, et al. In both this project and Ruffner and Bickley, the primary sampling unit was ATM tasks and not subjects. A total of 25 tasks were found which were common to the tasks rated as part of the Delphi project. Because the standard deviation of the Delphi ratings of these tasks ( $sd = 103.8$ ) approximated the mean ( $M = 129.95$ ) a common log transformation was performed on the data. This is not atypical of psychophysical data where there is no upper or lower anchor on estimates; consequently, all subsequent analyses of the Delphi data will employ a log transformation. The resultant  $M$  and  $sd$  were, respectively, 1.99; .33. For IP ratings of pilot performance, these were: ( $M = 4.34$ ;  $sd = .55$ ).

The resultant correlation between the two sets of ratings was highly significant ( $r = -.77$ ,  $df = 23$ ,  $p < .001$ ), indicating that estimates of ATM task difficulty did predict IP ratings of performance of nonproficient pilots on the same tasks. Roughly half of these aviators ( $n = 24$ ), returned for proficiency training the second year. The  $r$  between Delphi ratings and second year initial checkride IP ratings for this subgroup of pilots was also significant ( $r = -.73$ ,  $df = 18$ ,  $p < .005$ ). The degrees of freedom are fewer in this case because fewer ATM tasks are reported for the second year subsample.

SME estimates of time to perform were also considered a candidate index of task demand on the Delphi study. For this variable,  $M = 4.90$  minutes;  $sd = 2.83$ . Because of the large amount of variation in this data, a not uncommon occurrence for time estimates, a log transformation was performed, resulting in an  $M$  of .61 and  $sd$  of .28. Estimated time to perform a given ATM task

correlated moderately and significantly with estimated difficulty ( $r = .62$ ,  $df = 23$ ,  $p < .005$ ), and with IP ratings of performance ( $r = -.49$ ,  $df = 23$ ,  $p < .025$ ).

It may be informative to note that for the small subsample of aviators who returned the second year, the correlation between the Delphi estimates of time required to perform an ATM task and IP ratings of performance was highly significant for the second year initial checkride ( $r = -.59$ ,  $df = 18$ ,  $p < .005$ ), but not the first ( $r = -.26$ ). Recall that the latter correlation was significant when the whole sample was included.

Because time to perform and difficulty were correlated with one another and also with the criterion performance rating, partial correlations were computed for these variables. Partialling out the effects of time to perform, the correlation for difficulty on performance ratings was still highly significant ( $r = -.68$ ;  $p < .001$ ). The  $r$  for time to perform on IP ratings of performance, holding difficulty constant, was not significant ( $r = -.02$ ). SME estimates of task difficulty alone, irrespective of estimated time to perform, were a strong correlate of initial proficiency flight ratings of pilot performance.

Correlation with Ruffner and Bickley. The Delphi ratings of task difficulty for the 24 ATM tasks correlated significantly with IP performance ratings on the initial checkride ( $r = -.79$ ,  $df = 22$ ,  $p < .001$ ), and with final checkride performance ratings ( $r = -.62$ ,  $df = 22$ ,  $p < .005$ ). The correlation between initial and final checkride IP ratings was .90 ( $p < .001$ ).

Performance ratings for this research effort were made on a 12-point scale. Respective means and standard deviations were: 7.58, .64 (initial checkride); 7.78, .52 (final checkride). The time to perform estimates from Delphi yielded a mean of 5.99 minutes and standard deviation of 8.14. Because of this variation, a log transformation was performed, yielding a mean of .62 and a standard deviation of .33. For these data, the (log) Delphi ratings of time to perform the ATM tasks did not correlate significantly with IP ratings for initial ( $r = -.31$ ,  $df = 22$ ,  $p < .10$ ) or final ( $r = -.18$ ) checkrides.

Partialling out the effects of performance time, the first-order correlation between difficulty and IP ratings for the initial checkride was virtually unchanged ( $r = -.80$ ;  $p < .001$ ). The partial correlation for time to perform on the same criterion, controlling statistically for difficulty, changed to positive but was not significant ( $r = .37$ ;  $p < .10$ ).

Partial correlations were also computed, using final checkride scores as the criterion. The correlation between difficulty and IP performance ratings, holding constant time to perform, was significant ( $r = -.66$ ;  $p < .001$ ). The  $r$  of .32 between time to perform and the same criterion, was not significant.

Table 3 presents the transformed modified Delphi ratings for 20 ATM tasks which are common across all three projects.

Table 3

Log Modified Delphi Ratings of Difficulty and Time to Perform for ATM Tasks Common to Wick, et al. (1986) and Ruffner & Bickley (1985).

ATM Task Description	Log Delphi	
	Difficulty	Time (min)
Antitorque malfunction	2.66	.792
Climbs-Descents	1.65	.550
Confined area operations	2.30	.922
Deceleration-acceleration	2.00	.446
Engine failure (altitude)	2.04	.605
Engine failure (hover)	2.05	.513
Go-around	1.70	.290
Hover power check	1.60	.314
Hovering flight	1.60	.600
Hydraulic failure	2.16	.762
Landing from a hover	1.64	.270
Maximum performance takeoff	2.16	.516
Normal takeoff	1.95	.427
Pinnacle-ridgeline	2.31	.906
Steep approach	2.15	.706
Straight & level flight	1.53	.900
Standard autorotation	2.38	.948
Takeoff to a hover	1.60	.068
Traffic pattern flight	2.02	.957
Turns	1.70	.289

Delphi ratings of difficulty and accidents. In order to explore the application of the modified Delphi ratings of task difficulty to ATM tasks reported by ASMIS, 141 UH-60 accidents involving human error were examined. From the total number of accident report summaries, 99 usable cases, subsumed under 28 ATM tasks, were retrieved. These were cases where responsibility for the accident was attributed to the pilot, copilot, instructor pilot, or student pilot. The current research effort sought simply to match each ATM task description in the ASMIS to the modified Delphi rating for the same task.

An examination of Table 4 indicates that the most frequent task categories associated with accidents were those involving various phases of terrain flight ( $n = 21$ ), followed by phases of

landing (from a hover and roll-on;  $n=18$ ), and confined area operations ( $n=10$ ). It should be noted that although less demanding than most other accident-related ATM tasks, ground taxi accounts for a total of nine accidents.

The right-hand column of Table 4 lists 20 accidents that were Class A (loss of aircraft, fatality, or at least \$ .5 million). Note that for hard turns (evasive maneuvers) all accidents fell into Class A; for hovering flight, a task SMEs did not perceive as inordinately difficult, 66% of all accidents were class A.

A total of 25 mishaps involved night vision goggle (NVG) flight. A question quite pertinent to the present investigation is whether Class A and B accidents occur disproportionately under NVG conditions. A comparison of the relative frequencies showed that 28% (7) of the NVG accidents were class A or B vs. 26% (19) for non-NVG conditions. Thus, for the UH-60, it seems that the use or nonuse of NVGs has little to do with the severity of the accident.

Table 4

Aircrew Training Manual (ATM) tasks associated with UH-60 accidents.

ATM Task Title	Delphi	Freq.	Class A
Antitorque malfunction	400	1	
Circling approach	138	2	
Circling approach, terrain flight	164	1	
Confined area operations	200	9	1
Deceleration-acceleration	100	1	
Doppler navigation	154	2	
External load operations	240	7	1
Ground taxi	80	9	2
Evasive maneuvers (hard turns)	206	3	3
Hovering flight	40	6	4
Hydraulic malfunction	228	1	
Landing from a hover	93	13	
Landing from a hover, degraded AFCS	240	1	
Maximum performance takeoff	144	1	
Negotiate wire obstacles	180	2	
Normal takeoff	92	2	
Preflight inspection	118	1	
Roll on landing	160	4	
Single engine landing	172	1	1
Slope operations	150	1	1
Stabilator malfunction	90	1	
Terrain flight	130	14	5
Terrain flight approach	143	3	
Terrain flight takeoff	100	1	
Traffic pattern flight	102	3	1
Turns	50	1	
VMC approach	125	5	
Vertical IFR recovery procedures	212	3	1

One fundamental assumption of the present research effort was that high task demands, as expressed by the Delphi ratings, should be systematically related to the occurrence of accidents. The workload imposed by high task demands should make the occurrence of errors and consequently, accidents, more likely. The Delphi ratings of all 137 ATM tasks for the UH-60 showed an  $\bar{M}$  of 137.16 and an  $\underline{sd}$  of 101.00. Mean and standard deviation for the Delphi ratings of the subset of accident-related tasks ( $n=28$ ) were, respectively, 151.89; 72.69. For those remaining tasks that were not reported in conjunction with any accidents,  $\bar{M}= 119.67$ ;  $\underline{sd}= 84.23$ .

The reader should note that the standard deviation of this data set is high in relation to the mean. A log transformation was considered justified for this reason. The resultant means and standard deviations of the transformed data indicated that the transformation was successful. For all 137 tasks,  $\bar{M}= 1.99$ ;  $\underline{sd}=.32$ ; for the accident-related subset of 28 tasks,  $\bar{M}= 2.13$ ,  $\underline{sd}=.21$ ; for the non-accident-related tasks,  $\bar{M}= 1.95$ ,  $\underline{sd}=.33$ .

The Delphi ratings for accident and non-accident ATM tasks were contrasted via a  $t$ -test. The resulting  $t$  ratio ( $t= 2.92$ ,  $df= 135$ ,  $p< .01$ ; two-tailed test) was significant. In order to determine the degree of association between Delphi ratings and the accident vs. non-accident classification of the ATM tasks, a point-biserial correlation was computed. The resulting  $r_{pb}$  of .24 was significant ( $p< .05$ ).

One might argue that it is a fairer comparison to weight the tasks in Table 4 by their frequency of occurrence. This was done, yielding a respective (log) mean and standard deviation of 2.10; .20, which is almost identical to the result obtained without weighting.

#### DISCUSSION

In general, it appears that the secondary analyses of the data of both these research projects supported the hypothesis that the modified Delphi ratings of task difficulty would correlate negatively with IP ratings of pilot performance. This is consistent with the rationale underlying most notions of subjective indices of workload.

The Delphi performance time estimates for the same ATM tasks did not show such clear-cut results. In the case of the first project (Wick, et al.), they correlated significantly and negatively with ratings of performance for the entire sample as well as for the initial checkride of a 51% subsample that returned a year later; for Ruffner and Bickley, neither correlation with the first nor the second checkride was significant.

The partial correlation coefficients for difficulty and performance time estimates, computed for both research projects, indicate that the relationship between performance time and IP ratings may be more complex than originally supposed. For Wick, et al. it appears that the significant correlation between time and IP ratings was due primarily to the moderately high correlation between time to perform and difficulty. When difficulty is held constant, the correlation between time to perform and IP ratings becomes virtually zero. For Ruffner and Bickley, the zero-order correlations between time to perform and IP ratings were negative and nonsignificant. When the effects of difficulty were controlled statistically, however, these

correlations for both initial and final checkride became positive and approached significance.

This anomalous and intriguing finding is difficult to explain on a post hoc basis. One tentative explanation might be that some degree of skill decay is required before the time needed to perform a task covaries with difficulty. Recall that the Wick, et al. project consisted of reserve aviators who were much less proficient than those in the Ruffner and Bickley research effort. Thus, when skills are current, and most psychomotor tasks overlearned, the more difficult task may not take significantly longer to perform than one which is less difficult. The highly proficient aviator may even perform better on those tasks which require more time, simply because this allows for more practice.

### Correlations with IP Ratings

These intercorrelations confirm that the modified Delphi estimates have some validity in that they show that the more difficult a task is, the worse a pilot's performance on that task. This relationship was found to hold true whether or not the pilot was proficient. In general, more difficult tasks take longer to perform than less difficult tasks. The greater the difficulty of a task, the more performance can be expected to deteriorate with long periods of nonpractice. The latter findings seem hardly surprising if not obvious. What was somewhat surprising, however, was the magnitude of the correlation between the subjective Delphi estimates and IP ratings of pilot performance on the initial proficiency flight. It is true that the subject aircraft for both sets of ratings were different (UH-1 vs. UH-60); however, both are utility aircraft with essentially identical missions. The methods of ratings were also quite different (magnitude estimation vs. 7 and 12-point scales).

In short, it appears that the present results suggest that the methodology used in the Lofaro modified Delphi research yields valid weights by which the demands of ATM tasks can be assessed.

### Accident Prevention Usage

The derivation of these weights for aircraft like the UH-60 could provide an index of subjective workload and time demands, which could provide guidance for predicting "high-risk" phases of a mission where the pilot is likely to be overloaded, and where slips and mistakes are likely to occur. This could in turn provide a starting point for planning the management of workload through crew coordination, focusing initially on high-workload tasks which require more time-sharing than those which are less demanding.

The corollary finding that the more difficult ATM tasks are more likely to be reported by ASMIS as accident-related, than are those rated as less difficult, suggests a potentially useful means of singling out those problem tasks that are apt to be associated with mishaps. This in turn would suggest training countermeasures and training time priorities (such as increased practice time for problem tasks) which could result in greater proficiency and hence, lessen the probability of poor performance on these safety-critical tasks.

### Limitations

It is necessary to be aware of the pitfalls of this kind of post hoc, exploratory analysis. The chief difficulty is the fact that the data from the two ARI-sponsored projects on pilot proficiency are aggregate; the unit of analysis is mean IP ratings for whole groups of aviators rather than the ratings of individuals. In social science disciplines where post hoc, archival research is common, the use of data consisting of means or ranks is considered a potential source of bias which may possibly inflate the size of correlations so that they appear to be more significant than they really are, or appear significant when they, in fact, are not. Under the present circumstances, there was no way in which this problem could have been circumvented. It should suffice to state that the present results should be interpreted cautiously with this in mind.

Acknowledging these prior caveats, it would still seem that on the basis of their magnitude, the correlations obtained are a robust measure of the validity of the modified Delphi ratings. The replication of these correlations across two independent sets of checkride performance ratings bolsters this argument. Bearing in mind that these findings are the result of a secondary analysis of unrelated research projects, it would seem that the next step would be a direct predictive validation of the modified Delphi data against objective performance measures in the simulator. This in turn would allow investigators to determine if these subjective ratings of task difficulty actually do predict pilot performance.



## REFERENCES

- Casali, J.G. and Wierwille, W. W. (1983). A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load. Human Factors, 25, 623-642.
- Dalkey, N.C. (1969). The Delphi method. Rand Corporation Monograph (Whole, RM-5888).
- Glass, G. (1976). Primary, secondary and meta-analysis of research. Educational Research 5, 3-8.
- Gopher, D. and Braune, R. (1984). On the psychophysics of workload: Why bother with subjective measures? Human Factors, 26, 519-532.
- Hart, S.G. and Bortolussi, M. R. (1984). Pilot errors as a source of workload. Human Factors, 26, 545-556.
- Lofaro, R. J. (1985). Methodological modifications and considerations for a new small-scale Delphi paradigm. Unpublished manuscript, ARI Ft. Rucker Field Unit.
- Moray, N., Johanssen, J., Pew, R.D., Rasmussen, J., Sanders, A.F., & Wickens, C.D. (1979). Report of the experimental psychology group. In N. Moray (Ed.), Mental workload, its theory and measurement. New York: Plenum.
- Morris, N.N. and Rouse, W. B. (1985). An experimental approach to validating a theory of human error in complex systems. Proceedings of the Human Factors Society 29th Annual Meeting, 333-337.
- Ruffner, J. W. and Bickley, W. R. (1985). Validation of Aircrew Training Manual practice iteration requirements. ARI Technical Report 696, AD A 173 441.
- Stevens, S.S. (1971). Issues in psychophysical measurement. Psychological Review, 78, 426-450.
- Vidulich, M.A. and Tsang, P.S. (1985). Assessing subjective workload assessment: A comparison of SWAT and NASA bipolar methods. Proceedings of the Human Factors Society 29th Annual Meeting, Baltimore, 71-75.
- Wick, D. T., Millard, S.L. and Cross, K.D. (1986). Evaluation of a revised Individual Ready Reserve (IRR) Aviator Training Program: Final report. ARI Technical Report 697, AD A 173 811.

## TASK 1028: Perform VMC Approach.

CONDITIONS: In a UH-60 helicopter or a UH60FS with before landing check completed.

## STANDARDS:

1. Select a suitable landing area.
2. Establish the proper altitude to clear obstacles on final approach, and maintain altitude + or - 100 feet.
3. Establish entry airspeed + or - 10 KIAS.
4. Maintain a constant approach angle to clear obstacles.
5. Maintain ground track alignment with the landing direction with minimum drift.
6. Maintain apparent rate of closure, not to exceed the speed of a brisk walk.
7. Execute a smooth and controlled termination to a hover or to the ground.

## DESCRIPTION:

1. To a hover. Determine an approach angle which allows safe obstacle clearance while descending to the intended point of landing. Once the approach angle is intercepted (on base or final) adjust the collective as necessary to establish and maintain the angle. Maintain entry airspeed until apparent ground speed and rate of closure appear to be increasing. Progressively decrease the rate of descent and rate of closure until appropriate hover is established over the intended termination point. Maintain ground track alignment with the landing direction by maintaining the aircraft in trim above 50 ft AGL and aligning the aircraft with the landing direction below 50 ft AGL.

2. To the ground. Proceed as for an approach to a hover, except continue the descent to the ground. Make touchdown with minimum ground movement. After the landing gear contacts the ground, ensure the aircraft remains stable with all movement stopped. Smoothly reduce the collective to full-down position, and neutralize the pedals and cyclic.

NOTE 1: The decision to go-around should be made before descending below obstacles or decelerating below ETL.

NOTE 2: For training, recommended airspeed is 80 KIAS.

## APPENDIX A (Continued)

NOTE 3: Refer to FM 1-202 for procedures to reduce the hazards associated with the loss of visual references during the landing because of blowing snow or dust.

### NIGHT OR NVG CONSIDERATIONS:

#### 1. Night.

a. Altitude, apparent ground speed, and rate of closure are difficult to estimate at night. The rate of descent during the final 100 ft should be slightly slower than during the day to avoid abrupt attitude changes at low altitudes. After establishing the descent, reduce airspeed to approximately 50 KT until apparent ground speed and rate of closure appear to be increasing. Progressively decrease the rate of descent and forward speed until termination.

b. Be aware that surrounding terrain or vegetation may decrease contrast and cause a degradation of depth perception during the approach to the landing area. Before descending below obstacles, determine the need for artificial lighting.

#### 2. NVG. See TASK 2096.